How Do We Experience Crossmodal Correspondent Mulsemedia Content?

Alexandra Covaci, Estêvão B.Saleme, Gebremariam Mesfin, Nadia Hussain, Elahe Kani-Zabihi, Gheorghita Ghinea, *Member,IEEE*

***Abstract*—Sensory studies emerged as a significant influence upon Human Computer Interaction and traditional multimedia. Mulsemedia is an area that extends multimedia addressing issues of multisensorial response through the combination of at least three media, typically a non-traditional media with traditional audio-visual content. In this paper, we explore the concepts of Quality of Experience and crossmodal correspondences through a case study of different types of mulsemedia setups. The content is designed following principles of crossmodal correspondence between different sensory dimensions and delivered through olfactory, auditory and vibrotactile displays. The Quality of Experience is evaluated through both subjective (questionnaire) and objective means (eye gaze and heart rate). Results show that the auditory experience has an influence on the olfactory sensorial responses and lessens the perception of lingering odor. Heat maps of the eye gazes suggest that the crossmodality between olfactory and visual content leads to an increased visual attention on the factors of the employed crossmodal correspondence (e.g., color, brightness, shape).**

***Index Terms*—mulsemedia, QoE, crossmodal correspondence, heart rate, eye gaze.**

I. Introduction

THE Qualinet White Paper defines Quality of Experience (QoE) as "the degree of delight or annoyance of the user of an application or service" [5]. QoE is not a technical metric, but rather a concept that encapsulates all the elements related to a user's perception of a certain service and can be influenced by factors such as content, network, device, or context of use.

In a range of applications, content has started to go beyond the traditional audio visual dimensions. Over the last decade, researchers in multimedia, HCI or virtual reality have started to increasingly capitalize on touch, taste, and smell when designing tasks and interactions. As emphasized in [40], multisensory experience design has a promising potential on markets and society, leading to the creation of new products and experiences.

However, despite the recent advances and interest, there is no set of clear guidelines on how to create a multisensory content the users enjoy and benefit from. The reasons are multiple, from highly variable hardware platforms to human-centred problems related to human perception and preferences.

Mulsemedia (multiple sensorial media) can be seen as an extension of multimedia that represents media applications which go beyond engaging the traditional auditory and visual senses [15]. Mulsemedia gets closer to our experiences of everyday life by stimulating senses such as touch [12], smell [14] or taste [41], [42] with the aim to increase the user's QoE and to explore novel methods for interaction [6], [36].

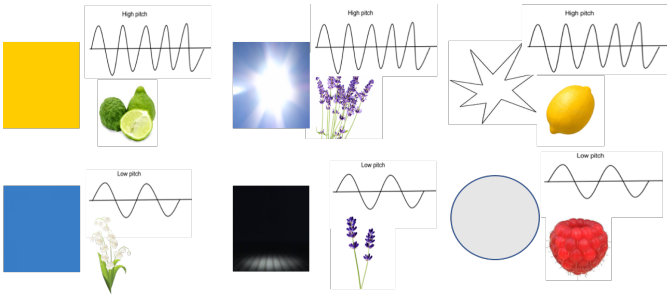
QoE is one of the important metrics to consider when building a system, and it constitutes an indicator of how well this system meets its targets. User QoE is traditionally assessed either through subjective methods such as questionnaires [34], [52], [58] or via objective metrics like electrodermal activity or heart rate [11]. QoE can be affected by different types of internal and external factors that should be handled from a holistic and perceptual point of view [56]. In [56], the author emphasizes the importance of visual attention when users are watching a video presentation, proposing to consider this attention mechanism in defining and assessing QoE.

User QoE in existent mulsemedia systems has been studied mostly from the perspective of inter-stream synchronization between different types of sensory data [2], [35], [39], [49], [58], [59], with only isolated instances looking at, for example, the impact of olfactory congruence on user QoE [16]. However, in order to enhance the quality of multisensory experiences, one needs more than just spatio-temporal integration; to this end, semantic and crossmodal correspondences play an important role.

In this paper we bridge the gap between studies of crossmodal correspondences and QoE by examining the effects of crossmodal congruence on engagement. We hypothesize that studying the influence of crossmodal correspondences on QoE, in a digital setup, could bring interesting insights related to content production and interaction. In this paper, we explore how users experience different types of multisensory content that is designed considering crossmodal correspondence principles, as described in the literature [45]. More precisely, we choose the dominant visual features of several videos and we add layers of auditory, olfactory and vibrotactile content that are crossmodally correspondent to these features. Based on these videos, we investigate the impact of the auditory and olfactory content on the users' QoE in a mulsemedia setup designed on principles of crossmodal correspondence. The evaluation of QoE in these cases is subjective (questionnaire) and objective - we analyze the gaze pattern to obtain insights on visual attention and the heart rate of participants to understand how the levels of excitement varied across conditions.

Case study. In recent years, Internet video streaming has become a dominant contributor to the global Internet traffic. Since video streaming services are bandwidth-hungry, it is challenging to maintain the QoE when bandwidth resources

A. Covaci is with University of Kent, UK
E. B. Saleme is with Federal University of Espírito Santo, Brazil.
E. Kani-Zabihi is University of West London, UK.
G. Mesfin, N. Hussain, and G. Ghinea are with Brunel University, UK.



(a) Color (b) Brightness (c) Shape

Figure 1: Olfactory and auditory crossmodal correspondences for different visual features (color, brightness, shape)

become scarce [24]. An interesting approach that uses deep neural networks for maximising the user QoE at the client side is presented in [55]. Given the demonstrated influence of mulsemedia on QoE, another way to maximise this metric at the client side could be through multisensory stimulation. However, the production of this type of content is done through annotation tools, in a tedious process.

We argue that this annotation process can be automatised through machine-learning methods that based on shapes, colours [27], [53] or other features of the video decide on multisensory content to stimulate a variety of senses [21], [22], [57]. If our hypothesis holds, this content can be created considering crossmodal correspondence principles. This approach could have meaningful applications in: a) increasing user QoE of different types of media by accommodating user's sensorimotor skills; b) providing non-verbal communication channels between interface designers and end users [23], [28]; c) building content for sensory substitution that could benefit people with sensory impairments (e.g., blind users [1], [19]); d) improving content recall through multisensory congruency in a variety of mulsemedia applications [32].

The remainder of this paper is organized as follows. Section II discusses related work focusing on different aspects of crossmodal correspondences, Section III describes the user study and the assessment methodology employed. Section IV presents the test results and analysis, whilst Section V concludes the paper and sets directions for future research.

II. Background

A. Olfactory-visual crossmodal correspondences research in psychology

Our senses do not operate in a sensory vacuum, and researchers have shown the existence of compatibility effects between stimuli in different sensory modalities. These mappings, called crossmodal correspondences, occur in a manner that is surprisingly consistent (e.g., pitch in audition and brightness in vision).

Crossmodal congruences play an important role in multisensory integration and they might occur because stimulus dimensions are correlated in nature and in the way we experience them, but also because of innate neural connections [31].

Crossmodal correspondences were documented between various pairs of sensory modalities such as: vision and touch [44], audition and touch [54], flavors and sounds [7], flavors and vision [13]. The focus of this paper are the associations between vision and olfaction and vision and audition, part of them being illustrated in Figure 1.

Presentation of visual stimuli can influence olfactory information processing and vice versa. Visual information affects the olfactory perception in many aspects such as intensity [26], pleasantness [43] and quality evaluations [47]. Although different visual stimuli have been analyzed in previous research, color was the predominantly used feature. In [26], more intense smells were associated with darker colors (Figure 1b). In [17], the authors present a study on color-odor linkages that showed consensus between lilial scent - blue color, cinnamon scent - red color, bergamot scent - yellow color, etc., part of these matches being illustrated in Figure 1a. Other studies focused on the shape - color correspondences and found that lemon and pepper odors are significantly associated with the angular shape, whereas the raspberry and vanilla odors were significantly associated with round shapes (see Figure 1c) [20]. Additionally, the same figure shows that these visual features also match certain audio characteristics predominantly related to pitch, as demonstrated in [31].

Crossmodal correspondences were also shown to influence people's performance under different experimental paradigms: direct crossmodal matching, faster classification tasks, faster simple detection tasks, Implicit Association Tests, spatial localization tasks, and perceptual discrimination tasks [45]. For instance, when participants were exposed to white or black visual stimuli, their speed of classification was faster when the color was accompanied by congruently pitched auditory stimuli [31]. This illustrates the potential of the correspondence between sound frequency and color brightness in performing cognitive tasks. This potential was investigated in [18], where the authors showed that high-frequency sounds guide the visual attention toward light-colored objects, while low-frequency sounds guide it toward dark-colored objects.

The practical relevance of this area started to attract the attention of food sectors, marketers and advertisers who became interested in how they can convey information about the fragrance/flavor/taste of their products by making use of the matches between the attributes of stimuli in different sensory dimensions [46].

B. Olfactory-visual crossmodal correspondences applications outside psychology

Because all our senses interact to influence our experiences, different types of sensory cues can be used to guide or modify sensory expectations, search and augmentation. As shown in the previous section, crossmodal correspondences between different sensory dimensions show an interesting practical potential. However, the work on this outside the field of cognitive sciences is limited, notwithstanding the fact that Understanding cognitive processes across all sensory modalities could play an important role in developing efficient digital multimodal interfaces that consider all the subtleties involved in human perception [50].

Crossmodal correspondences support comprehension and retention of information through the accommodation of users' sensorimotor skills. Thus, their application in different contexts, such as interaction design, computer graphics, information retention, QoE could bring interesting insights.

In computer graphics, displaying crossmodaly-linked content has been shown to distract viewers from correctly identifying animation quality [4]. Promising applications for crossmodal correspondences in interaction design were presented in [10]. The authors explored the efficiency of olfaction in introducing a new semantic layer in HCI, in a study where they analyzed different mappings between driving-relevant messages and scents. Strong associations were found between the "Slow down" message and the scent of lemon, the "Fill gas" message and the scent of peppermint and between "Passing by a point of interest" message and the scent of rose. This confirmed the hypothesis of the authors that using crossmodally congruent olfactory information can transfer specific visual information to the user. In [48], the authors investigated the effects of employing crossmodal correspondences between haptic and audio output on augmenting focus. They used these principles in the design of "atmoSphere", a sphere that provides haptic feedback designed to augmented focus during mindfulness training by guiding the users into a particular rhythm of breathing. Participants to this study indicated that they had an entertaining experience, however this is dependent on the type of sound and haptic feedback (foot steps and rain drops seemed to work best).

C. Predictions and overview

Mulsemedia contributes directly and indirectly to the user perceived QoE [2], [58] by enriching the levels of enjoyment of applications or by masking a decreased quality of the audiovisual stream or the synchronization skews. However, although adding sensory modalities has shown general positive results on improving the user experience, crossmodal correspondences have rarely been considered in the design of content for mulsemedia systems.

In a practical scenario that involves mulsemedia services (e.g., watching a movie while experiencing a certain scent), QoE can also be formulated as the acceptable combination of different sensory inputs. As such, in [56], the authors proposed the integration of visual attention models in QoE assessment. In [11], QoE was evaluated through a combination of subjective (questionnaire) and objective methods (heart rate and electrodermal activity) as an indicator of the physiological arousal. The same objective metrics were used also in [25] in the assessment of immersive experiences in augmented reality applications.

User experience in mulsemedia represents a promising setup to investigate crossmodal correspondences outside of traditional multimedia systems. In this paper, we propose a mulsemedia setup, where visual content is delivered with different types of auditory, olfactory and vibrotactile content (crossmodally correspondent or not). Our aim is to investigate whether crossmodally matched content leads to an increased QoE. We evaluate this by compiling a set of subjective (questionnaire) and objective methods (heart rate and gaze patterns - visual attention) for the assessment of the user experience.

III. User study

The experiments we designed focus on a mulsemedia setup with visual, auditory, olfactory and vibrotactile content. We consider the visual stimulus serving as an attended sensory input. The visual content consists of six videos characterized by certain dominant visual features: color (blue, yellow), brightness (low, high), and shape (round, angular). The olfactory and auditory content is chosen so that they respect or not principles of crossmodal correspondence. The frequency of the auditory content serves also as input to a haptic vest with vibration motors, that creates vibrotactile effects. The choice of this vibrotactile display was made because of the reported users' increased emotional response to haptic-enhanced media [51].

A. Participants

We recruited 24 participants (12 females, 12 males) for a mixed study where olfactory, auditory and vibrotactile information were manipulated between subjects (across replications), while the visual content was varied within subjects. They were aged between 18 and 41 years old and hailed from diverse nationalities and educational backgrounds (undergraduate and postgraduate students as well as academic staff). All participants spoke English and self-reported as being computer literate. They were randomly assigned to one of four different groups (Table I) taking into account age and biological differences between genders.

B. Experimental apparatus

The experiments took place in a laboratory environment with good ventilation, thus avoiding the problem of smell mixing. The setup we used for all the experiments is presented in Figure 2.

The olfactory content was provided by Exhalia SBi4, an olfactory device considered by previous research more reliable and more robust than other existing devices [37]. The olfactory display was placed at 0.5m in front of the participant, allowing her/him to detect the smell in 2.7-3.2s [36]. SBi4 can store up to four interchangeable scent cartridges at a time. Scent is emitted while air is blown by the built-in-fans through cartridges that contain scented polymer. To prevent the mixing of scents, in our experiments we used a single cartridge. The olfactory content is synchronized with the visual content by a program that uses Exhalia's Java-based SDK.

An Eye Tribe eye tracker controlled by a custom written Java code was employed to record eye-gaze patterns on a Windows 10 Laptop with 8GB RAM powered by an IntelCore i5 processor. The viewing screen was placed between 45-75 cm from the eyes of the participants, as this was the recommended distance for Eye Tribe calibration[1]. We chose to use the EyeTribe eye tracker because of previous reports that showed its accuracy in studies on gaze points and fixations [8].

[1] http://theeyetribe.com

Group	Content		
	Olfactory	Auditory	Vibrotactile
G1	Congruent with the visual content: V1 - Lilial, V2 - Bergamot [29]; V3 - Clear lavender, V4 - Lavender [26]; V5 - Lemon, V6 - Raspberry [20]	Original audio	Auto-generated
G2	Rosemary	Congruent with the visual content: V1 - Low pitch, V2 - High pitch, V3 - Low pitch, V4 - High pitch, V5 - High pitch, V6 - Low pitch	Auto-generated
G3	Congruent with the visual content: V1 - Lilial, V2 - Bergamot; V3 - Clear lavender, V4 - Lavender; V5 - Lemon, V6 - Raspberry	Disabled	Disabled
G4	Rosemary	Disabled	Disabled

Table I: Stimuli assortments for the four experimental groups

The videos were displayed on the computer monitor with a resolution of 1366x768 pixels, with a viewing area of 1000x700 pixels in the center of the screen.

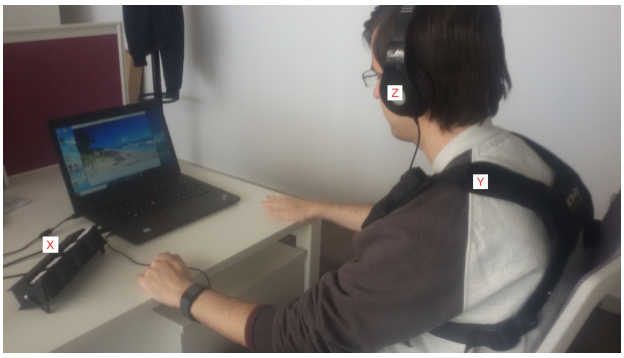


Figure 2: Experimental setup for the visual, audio, olfactory, haptic media display system (X - Exhalia SBi4-radio scent emitter, Y - KOR-FX Haptic Vest, Z - Headphones).

Participants sat in a chair without armrests facing the screen. All participants wore i-shine[2] headphones, a vibrotactile KOR-FX[3] gaming vest, and a Mio Link heart rate wristband[4]. To facilitate the vibrotactile experience, we chose the KOR-FX gaming vest that utilizes 4DFX based auditory-haptic signals to enable haptic feedback to the upper chest and shoulder regions. The vest is wirelessly connected to a control box meant to accept the standard sound output of the sound card of a computer. This type of devices deliver additional information about environmental factors while immersing users in the experience [33].

C. Mulsemedia content

As illustrated in Figure 3, the six videos used in our experiments were selected based on their dominant visual features such as color, brightness and angularity of objects. The videos were shot using a static camera, thus similar to a timelapse, the difference between frames was not significant. Hence, the areas of the snapshots presented in the Figure 3 and the scenes depicted remained the same over the duration of the videos. Four variants (described in Table I) were created from different types of olfactory, auditory and vibrotactile content that accompanied the videos.

The videos used in our experiment were 120 seconds long. The auditory content was adjusted for G2 to a frequency of 328Hz (high pitch condition) and 41Hz (low pitch condition). When present, the vibrotactile content was provided for the whole length of the video and was derived automatically from the audio content. The olfactory content consisted of seven scents: rosemary, bergamot, lilial, clear lavender (low intensity), lavender (high intensity), lemon and raspberry. These scents were delivered over a 60s interval, in the middle of the video (from second 30 to second 90). For this, a software framework has been developed to control the presentation of olfactory data and video. This allowed us to diminish the presence of any lingering scents. Additionally, in between the videos, while participants were filling in the questionnaire, the windows were opened, allowing new tests to take place in neutral conditions.

D. Procedure

Pre-experiment study - audio pitch. Before the experiments, we carried out a small pilot study with two participants, to get feedback on their thoughts and experiences while trying our system. This was aimed to inform us about any disturbing or distracting factors. Since they reported that the high pitch audio was distracting, we lowered its volume to increase their comfort during the experiment.

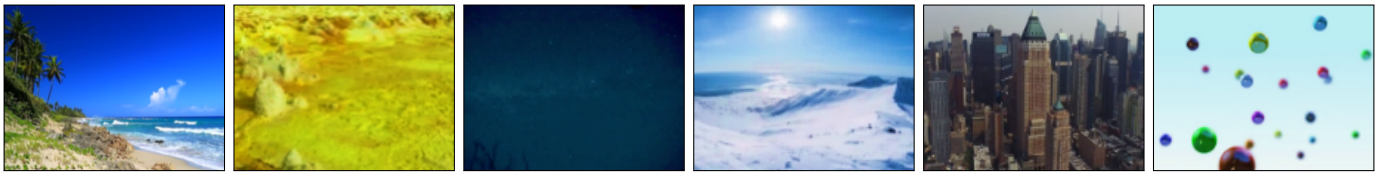
Conditions. Participants were randomly divided in four groups of six each (described in Table I) and watched the six videos in a random order. For G2 and G4, we used rosemary scent because of its demonstrated benefits on increasing alertness in tasks [9]. For G1 and G3, scents were selected based on olfactory-visual crossmodal principles.

Collected data.

Participants completed a subjective questionnaire consisting of eight questions at the end of each video (Table II). Each question was answered on a 5-item Likert scale, anchored at one end with ”Strongly Disagree” and with ”Strongly Agree” at the other.

Additionally, we recorded gaze data and heart rate for an objective evaluation of QoE.

[2]https://www.ishine-trade.com/Headphones-Earphones
[3]http://korfx.com/products
[4]https://www.mioglobal.com/



(a) **V1.** Color: Blue (b) **V2.** Color: Yellow (c) **V3.** Brightness: Low (d) **V4.** Brightness: High (e) **V5.** Shapes: Angular (f) **V6.** Shapes: Round

Figure 3: Snapshots from the videos used in the experiment and their corresponding dominant visual cue.

Questions
Q1: The smell was relevant to the video clip I was watching;
Q2: The smell came across strong;
Q3: The smell was distracting;
Q4: The smell was consistent with the video clip when released;
Q5: The smell was annoying;
Q6: The smell faded away slowly after watching the video clip;
Q7: The smell enhanced my viewing experience;
Q8: Overall, I enjoyed the multisensorial experience.

Table II: QoE questionnaire

IV. Results

In this section we present the results from our experiments in order to establish the influence of different types of auditory and olfactory content on the reported user QoE.

More precisely, we are interested to answer whether:

1) Does auditory content increase the QoE in setups where visual and olfactory content are crossmodally correspondent?
2) Does crossmodally congruent auditory content improve the QoE in the presence of rosemary odor?
3) Does crossmodally correspondent olfactory content increase QoE?
4) What is the influence of different types of content on user's gaze behavior?
5) What is the influence of different type of content on the heart rate as an indicator of QoE?

These shall now be looked at in turn.

A. Does auditory content increase the QoE in setups where visual and olfactory content are crossmodally correspondent?

To determine whether the presence of audio is important in the evaluation of the QoE of crossmodally matched olfactory and visual content, an independent samples t-test was performed on answers reported by Group 1 and Group 3. The p-Value was computed using SPSS for Windows (statistical presenting system software version 25.0) and $p < 0.05$ was considered to be statistically significant. Results indicate that there is a statistically significant difference in favour of the presence of the audio for **Q1**: t(70) = 3.463, p = 0.001; and **Q5**: t(70) = 2.875, p = 0.005. Significant values were obtained also for **Q6**: t(70) = -3.081, p = 0.003 and **Q8**: t(70) = -2.117, p =0.038. Mean values for the responses are showed in Figure 4.

This indicates that the presence of audio increased the perceived relevance of the olfactory content (Q1). However, when audio was enabled, the olfactory content was more annoying (Q5), but faded away faster between videos (Q6). This indicates that in the presence of the audio, the users were attending to vision, rather than to olfaction, thus they were less sensitive to the lingering of the scent. However, when it comes to the general evaluation of the multisensory experience, results showed that users enjoyed more the videos when the audio was disabled. This seems to indicate that the interplay between olfactory and auditory content does not have a positive impact on the enjoyment of the experience. Moreover, it shows the importance of using scent effects in multimedia and that the usage of audio in conjunction with (crossmodally matched) scents is detrimental to QoE - as opposed to the usually encountered situation whereby audio has primacy over video when only audio-visual content is employed. It is important to note that crossmodal correspondences have been studied almost exclusively in a laboratory setting with simple cues. However, our results are in line with observations in [32] stating that more complex conditions require a more careful design of crossmodal support.

In [3], the authors showed that for semantically congruent olfactory and visual content, the presence of audio content has a positive role on the users' perceived QoE (relevance, enjoyment, sense of reality). The results we obtained in this section show that for crossmodally congruent visual and olfactory content, the audio component does not contribute to the enhancement of enjoyment, but increases the sense of relevance of the odor. Although audio seems to make crossmodally correspondent odor more relevant, it also affects its hedonic dimension, users perceiving it more annoying. Our results show that when dealing with crossmodally congruent contents less is more and audio should be disabled to trigger an increase in enjoyment.

B. Does crossmodally congruent auditory content improve the QoE in the presence of rosemary odor?

In order to establish the effect of the crossmodally congruent auditory content on the reported QoE, we performed an independent samples t-test on the outcomes of Group 2 and Group 4. P values < 0.05 were considered significant. Results were statistically significant for **Q1**: t(70) = 4.533, $p < 0.001$; **Q3**: t(69) = 3.605, p = 0.001; **Q5**: t(69) = 4.412, $p < 0.001$; **Q6**: t(70) = -3.146, p = 0.001; and **Q8**: t(68) = -4.792, $p < 0.001$. Mean values are presented in Figure 5.

These results show that the presence of the audio content determined users to consider the displayed olfactory content as more relevant to the visual content (Q1). This seems to be consistent with the results reported in the previous subsection

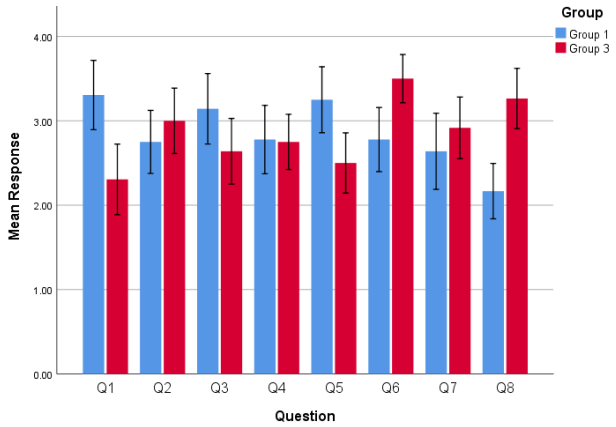


Figure 4: Responses of Group 1 and Group 3 for the QoE questionnaire

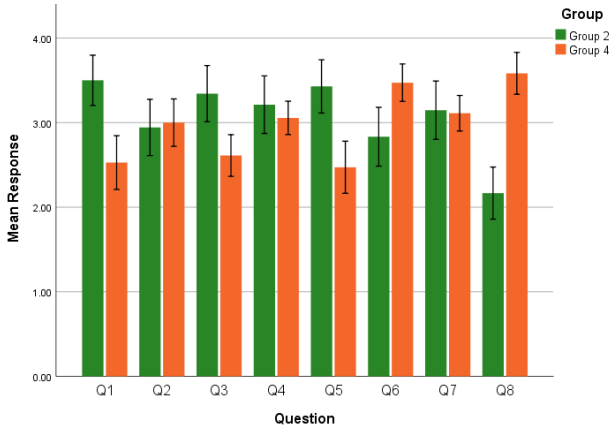


Figure 5: Responses of Group 2 and Group 4 for the QoE questionnaire

for the same question, where participants found the rosemary smell more distracting in the presence of audio than when the audio was disabled (Q3). However, the rosemary scent was perceived as less annoying in the presence of crossmodal correspondent auditory content (Q5). The fading time for rosemary smell was perceived shorter in the presence of audio content (Q6). This is consistent with the previous results. Participants evaluated better the overall QoE in the absence of audio content (Q8).

These results illustrate that crossmodally correspondent auditory content does not seem to have a positive effect on improving the overall mulsemedia experience. This shows that using the crossmodal congruency approach does not have the same positive results on the user QoE as traditional audiovideo setups, where audio and video are chosen based on other types of congruences (e.g., temporal, contextual). One possible explanation might lie in the observation that, in the pre-experiment study, users reported that the high audio pitch is distracting. Although we subsequently lowered the volume in the study proper, this might have still disturbed the users. Nonetheless, it is fair to say that a mixed picture emerges, for crossmodally correspondent auditory content, whilst heightening users' annoyance and distraction associated with the emitted scent, seems to also increase its perceived sense of relevance.

C. Does crossmodally correspondent olfactory content increase QoE?

To determine whether the type of olfactory content (crossmodally correspondent with visual content or rosemary) has an influence on the reported user QoE, we performed an independent samples t-test on the answers of Group 3 and Group 4. The level of significance was set at $p < 0.05$. Results did not show significant statistical differences between the two groups for none of the questions (mean values for the responses in Figure 6).

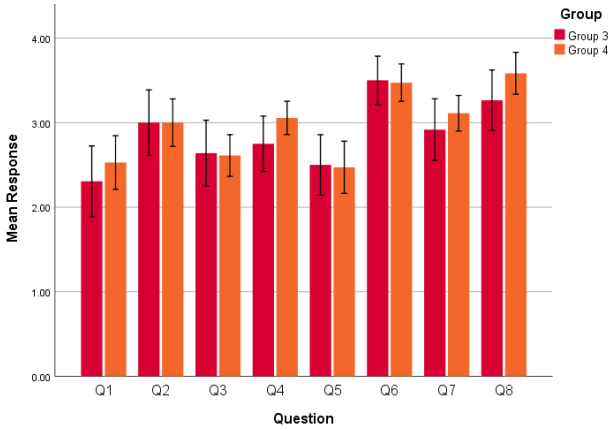


Figure 6: Responses of Group 3 and Group 4 for the QoE questionnaire

Studies on the QoE of setups where olfactory content was present are usually considering semantically congruent odors and synchronization aspects [2], [38]. Here we show that the nature of olfactory content (rosemary vs. crossmodally correspondent with visual features) does not seem to significantly affect the reported QoE when it comes to enjoyment and relevance. This finding echoes that of [16], which obtained a similar profile of results for when semantically incongruent vs. semantically congruent odors were used. Further investigation needs to assess whether other aspects of the experience are affected by the exposure to content designed based on crossmodal correspondence principles.

D. What is the influence of different types of content on user's gaze behavior?

Our goal was to understand how different types of mulsemedia content will impact on the attention of its recipients. Particularly, we wanted to investigate how crossmodal correspondences influence visual attention. Will the crossmodally matched additional content (olfactory, auditory, vibrotactile) guide the gaze of the user towards a certain type of visual feature? Will the users explore more or will they focus on certain areas of the screen in different experimental conditions?

To have a basis for the comparison, we used the eye gaze data recorded by the Eye Tribe and we plotted the heat maps for the visualization of the videos. We compared these results

with the ones provided by EyeQuant[5], a design assistant tool that uses a combination of leading neuroscience research and powerful AI to predict in real time how users will engage and react to any design. Attention maps using the EyeQuant tools are presented in Figure 7, where warm areas are predicted to be more visible.

For Group 1, the visual attention behavior for all the six videos is displayed in Figure 8. We can observe that when we deliver congruent olfactory and visual content, the gaze patterns differ from the ones estimated by EyeQuant. When they experienced lilial or bergamot odor, participants explored more the blue (Figure 8a) or yellow areas (Figure 8b). For V3, the eye gazes seem focused on the darker part of the video, where there is an agglomeration of branches (Figure 8c). A different gaze pattern is presented also for V4, where surprisingly, the focus is not on the brightest area of the video. Unlike the other videos, V5 is dynamic, so the camera perspective changes very often. Thus, it is hard to draw a conclusion about the effects of the lemon odor on the gaze of the users. While watching V6, participants focused on the incoming balls close to the center of the screen.

For Group 2 (where the users experienced auditory content crossmodaly congruent with the visual content while rosemary scent was emitted), visual attention seems to follow a different pattern indicating that pitch shifts the visual attention in a different way than odor. In V1 (Figure 9a), users are more focused on the beach and not on the blue waves or sky. In V2, they explore less the landscape, while in V3, their focus seems wider and more central than in case of Group 1. Another interesting difference is that they explore more the video with angular buildings. Overall, Group 2 shows similar results to the ones predicted by EyeQuest and this seems to indicate that this experimental condition does not bring significant changes when compared to a pure condition, where users are exposed only to visual content. Similar observations can be made also when for Group 3 (Figure 10), where the gaze patterns resemble the ones in Group 2 with the exception of Figure 10e, where participants explored less. Group 4 (Figure 11) was exposed to an experimental assortment where the olfactory content was rosemary and the auditory and vibrotactile contents were disabled. This time, the center of focus seems shifted for V1 (towards the sea), while for the other videos, the exploration region is broader.

Prior studies have indicated a shift in visual attention as an effect of the crossmodal correspondence between the visual and the auditory feedback [18]. Our observations showed that similar changes in the visual attention can be obtained also when using scent. Lilial odor automatically guides participants' attention toward blue-colored objects, and bergamot odor automatically guides participants' attention toward yellow-colored objects (Group 1 and Group 3). The influence of olfactory content matched with brightness or angularity visual features on the gaze patterns was not obvious. In our study, we did not observe a significant effect of the pitch on the gaze patterns: the behavior of Group 2 is similar to the behavior of Group 4.

Although we can not draw strong conclusions based on the gaze patterns of the participants, we observe that when the olfactory content is crossmodally congruent with the visual content (Group 1 and Group 3), the visual attention of the users seems shifted towards the correspondent visual feature (e.g., exploration and focus on the blue sky for V1; wider exploration area for the round shapes (more balls) for V6). Overall, it seems that visual attention tasks could benefit from the presence of additional content that matches the dimensions meant to be attended.

E. What is the influence of different type of content on the heart rate as an indicator of QoE?

An ANOVA with 95% confidence level was conducted to compare mean values for heart rate in all the groups with statistically significant results between all groups ($p < 0.000$). This revealed that the heart rate values differed significantly across conditions. Figure 12 presents the assessors average heart rate during the experience of the different experimental setups.

Differences in heart rates between the 4 groups of our study could be attributed to different moods experienced by the users when exposed to different sensory content. The highest average heart rate was registered for Group 1, where olfactory content was crossmodally correspondent, whereas the lowest occurs for Group 2 where auditory content was crossmodally congruent. This shows that the latter crossmodal correspondence induced a more relaxed mood in participants; in contrast, the former type of crossmodal congruence seems to lead to inreased levels of participant stress, as evidenced through their heart rates.

V. Conclusion

One aim of this article is to suggest that crossmodal correspondences are undervalued in the design of mulsemedia content. Mulsemedia is by definition an area rich in sensorial experiences. Whilst some insights into designing mulsemedia content have been progressed (most based on semnatically

[5]http://www.eyequant.com/

(a) **V1.** (b) **V2.** (c) **V3.** (d) **V4.** (e) **V5** (f) **V6.**

Figure 7: Attention maps for videos V1-V6 generated with EyeQuant

(a) **V1.** O: Lilial (b) **V2.** O: Bergamot (c) **V3.** O: Clear Lavender (d) **V4.** O: Lavender (e) **V5** O: Lemon (f) **V6.** O: Raspberry

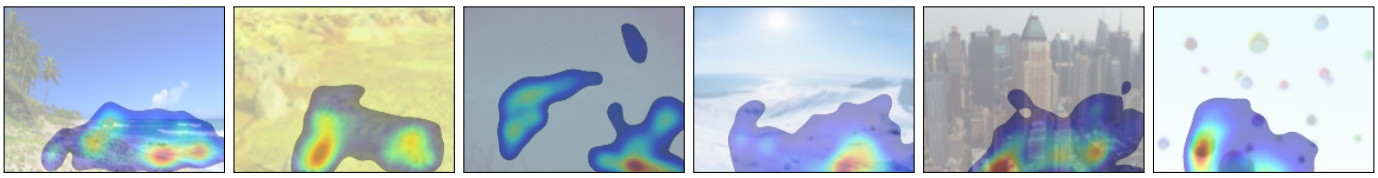
Figure 8: Gaze map for **Group 1**. Olfactory content (O): crossmodal. Auditory content - original audio. Vibrotactile content: generated from audio.

(a) **V1.** P: Low (b) **V2.** P: High (c) **V3.** P: Low (d) **V4.** P: High (e) **V5.** P: High (f) **V6.** P: Low

Figure 9: Gaze map for **Group 2**. Olfactory content: Rosemary. Auditory content - pitch (P): crossmodal. Vibrotactile content: generated from audio.

(a) **V1.** O: Lilial (b) **V2.** O: Bergamot (c) **V3.** O: Clear Lavender (d) **V4.** O: Lavender (e) **V5.** O: Lemon (f) **V6.** O: Raspberry

Figure 10: Gaze map for **Group 3**. Only olfactory content (O): crossmodal. Auditory and vibrotactile content: disabled.



(a) **V1.** O: Rosemary. P, VT: Off (b) **V2.** O: Rosemary. P, VT: Off (c) **V3.** O: Rosemary. P, VT: Off (d) **V4.** O: Rosemary. P, VT: Off (e) **V5.** O: Rosemary. P, VT: Off (f) **V6.** O: Rosemary. P, VT: Off

Figure 11: Gaze map for **Group 4**. Only olfactory content (O): Rosemary. Auditory - pitch (P) and vibrotactile (VT) content: disabled.

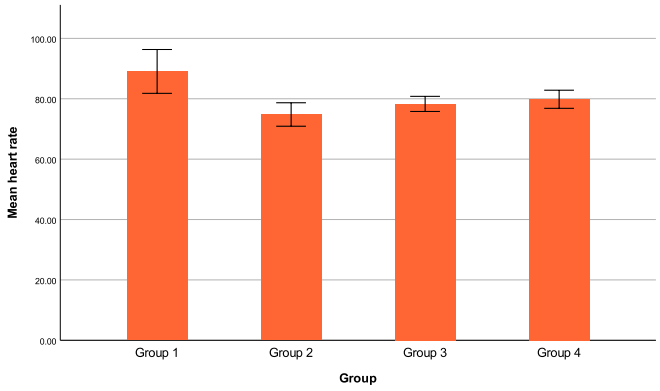


Figure 12: Heart rate mean across the four groups

congruent content), the relevance of crossmodal correspondences in this process has been neglected.

This paper proposes an alternative exploration to traditional studies on QoE by considering heart rate and gaze behavior in the evaluation. The findings of our study suggest that when crossmodal principles are considered in the design of mulsemedia systems, the content is perceived differently from when we use semantic congruence or other principles (e.g. rosemary odor is usually chosen in similar setups because it increases alertness in tasks). Crossmodally congruent olfactory content leads to a increased focus on the visual corresponding features, especially for the color and to an increased heart rate. However, when it comes to the reported QoE, it does not seem to create a significant impact. The effectiveness of crossmodally congruent audio content has yet to be further explored since our results show mixed aspects when it comes

to relevance and annoyance.

One of the limitations of the study is the relatively small number of participants that makes it unclear how these findings would generalise in other setups. Second, using specific videos makes it difficult t o p redict h ow u sers w ill e xperience other crossmodally correspondent setups. Nonetheless, our findings raise important questions that require further investigation related to how to inform better the design by engaging crossmodal interaction.

As future work, we would like to investigate the effect of crossmodally correspondent odors when performing tasks based on promising results from [30] where olfactory notifications were shown to improve users' confidence a nd performance. We believe that by using crossmodally correspondent scent, this learning process (that implied training participants to recognize odours) could be made more effective. Moreover, by combining EEG and eye movements we can achieve a better understanding of the way observers are engaging with the media and hence, an estimate of the effectiveness of a crossmodal solution and of the users' perceived quality. This will be the subject of further work together with a comprehensive study on the impact of human factors on the perceived quality in crossmodally correspondent setups. Finally, we are interested in exploring how the employment of crossmodal principles could benefit t he e ngagement o f people with a variety of sensory capabilities.

ACKNOWLEDGMENT

This paper was partially funded by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 688503. This study was also part financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 88881.187844/2018-01. Estêvão B. Saleme also thankfully acknowledges support from the Federal Institute of Espírito Santo.

REFERENCES

[1] Sami Abboud, Shlomi Hanassy, Shelly Levy-Tzedek, Shachar Maidenbaum, and Amir Amedi. Eyemusic: Introducing a visual colorful experience for the blind using auditory sensory substitution. *Restorative neurology and neuroscience*, 32(2):247–257, 2014.

[2] Oluwakemi A Ademoye and Gheorghita Ghinea. Synchronization of olfaction-enhanced multimedia. *IEEE Transactions on Multimedia*, 11(3):561–565, 2009.

[3] Oluwakemi A. Ademoye, Niall Murray, Gabriel-Miro Muntean, and Gheorghita Ghinea. Audio masking effect on inter-component skews in olfaction-enhanced multimedia presentations. *ACM Trans. Multimedia Comput. Commun. Appl.*, 12(4):51:1–51:14, August 2016.

[4] Belma R. Brkic, Alan Chalmers, Kevin Boulanger, Sumanta Pattanaik, and James Covington. Cross-modal affects of smell on the real-time rendering of grass. In *Proceedings of the 25th Spring Conference on Computer Graphics*, SCCG '09, pages 161–166, New York, NY, USA, 2009. ACM.

[5] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al. Qualinet white paper on definitions of quality of experience. 2013.

[6] Alexandra Covaci, Longhao Zou, Irina Tal, Gabriel-Miro Muntean, and Gheorghita Ghinea. Is multimedia multisensorial? - a review of mulsemedia systems. *ACM Comput. Surv.*, 51(5):91:1–91:35, September 2018.

[7] Anne-Sylvie Crisinel and Charles Spence. Implicit association between basic tastes and pitch. *Neuroscience letters*, 464(1):39–42, 2009.

[8] E. Dalmaijer. Is the low-cost eyetribe eye tracker any good for research? *PeerJ PrePrints*, 2014.

[9] Miguel A. Diego, Nancy Aaron Jones, Tiffany Field, Maria Hernandez-reif, Saul Schanberg, Cynthia Kuhn, Mary Galamaga, Virginia McAdam, and Robert Galamaga. Aromatherapy positively affects mood, eeg patterns of alertness and math computations. *International Journal of Neuroscience*, 96(3-4):217–224, 1998. PMID: 10069621.

[10] Dmitrijs Dmitrenko, Emanuela Maggioni, Chi Thanh Vi, and Marianna Obrist. What did i sniff? mapping scents onto driving-related messages. In *AutomotiveUI'17 Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 154–163. ACM, 2017.

[11] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray. An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2016.

[12] M. Eid, J. Cha, and A. El Saddik. Hugme: A haptic videoconferencing system for interpersonal communication. In *Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008. VECIMS 2008. IEEE Conference on*, pages 5–9. IEEE, 2008.

[13] David Gal, S Christian Wheeler, and Baba Shiv. Cross-modal influences on gustatory perception. 2007.

[14] G. Ghinea and O. Ademoye. The sweet smell of success: Enhancing multimedia applications with olfaction. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 8(1):2, 2012.

[15] George Ghinea, Frederic Andres, Stephen Gulliver, et al. *Multiple sensorial media advances and applications: New developments in MulSeMedia*. IGI Global, 2011.

[16] Gheorghita Ghinea and Oluwakemi Ademoye. User perception of media content association in olfaction-enhanced multimedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 8(4):52, 2012.

[17] Avery N Gilbert, Robyn Martin, and Sarah E Kemp. Cross-modal correspondence between vision and olfaction: the color of smells. *The American journal of psychology*, pages 335–351, 1996.

[18] Henrik Hagtvedt and S. Adam Brasel. Cross-modal communication: Sound frequency influences consumer responses to color lightness. *Journal of Marketing Research*, 53(4):551–562, 2016.

[19] Giles Hamilton-Fletcher, Marianna Obrist, Phil Watten, Michele Mengucci, and Jamie Ward. "i always wanted to see the night sky": Blind user preferences for sensory substitution devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2162–2174, New York, NY, USA, 2016. ACM.

[20] Grant Hanson-Vaux, Anne-Sylvie Crisinel, and Charles Spence. Smelling shapes: Crossmodal correspondences between odors and shapes. *Chemical senses*, 38(2):161–166, 2012.

[21] Di Hu, Xuelong Li, et al. Temporal multimodal learning in audiovisual speech recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3574–3582, 2016.

[22] Di Hu, Xiaoqiang Lu, and Xuelong Li. Multimodal learning via exploring deep semantic similarity. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 342–346. ACM, 2016.

[23] Olivia Jezler, Elia Gatti, Marco Gilardi, and Marianna Obrist. Scented material: Changing features of physical creations based on odors. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 1677–1683, New York, NY, USA, 2016. ACM.

[24] Junchen Jiang, Vyas Sekar, and Hui Zhang. Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive. *IEEE/ACM Transactions on Networking (ToN)*, 22(1):326–340, 2014.

[25] C. Keighrey, R. Flynn, S. Murray, and N. Murray. A qoe evaluation of immersive augmented and virtual reality speech amp;amp; language assessment applications. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, May 2017.

[26] Sarah E Kemp and Avery N Gilbert. Odor intensity and color lightness are correlated sensory dimensions. *The American journal of psychology*, 110(1):35, 1997.

[27] Xuelong Li, Aihong Yuan, and Xiaoqiang Lu. Multi-modal gated recurrent units for image description. *Multimedia Tools and Applications*, 77(22):29847–29869, 2018.

[28] X. Lu, B. Wang, X. Zheng, and X. Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, April 2018.

[29] M Luisa Dematte, Daniel Sanabria, and Charles Spence. Cross-modal associations between odors and colors. *Chemical Senses*, 31(6):531–538, 2006.

[30] Emanuela Maggioni, Robert Cobden, Dmitrijs Dmitrenko, and Marianna Obrist. Smell-o-message: Integration of olfactory notifications into a messaging application to improve users' performance. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 45–54, New York, NY, USA, 2018. ACM.

[31] Lawrence E Marks. On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13(3):384, 1987.

[32] Oussama Metatla, Nuno N. Correia, Fiore Martin, Nick Bryan-Kinns, and Tony Stockman. Tap the shapetones: Exploring the effects of crossmodal congruence in an audio-visual interface. In *CHI*, 2016.

[33] Gene Munster, Travis Jakel, Doug Clinton, and Erinn Murphy. Next mega tech theme is virtual reality. *gene*, 612:303–6452, 2015.

[34] N. Murray, B. Lee, Y. Qiao, and G. M. Muntean. The influence of human factors on olfaction based mulsemedia quality of experience. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2016.

[35] N. Murray, G. M. Muntean, Y. Qiao, S. Brennan, and B. Lee. Modeling user quality of experience of olfaction-enhanced multimedia. *IEEE Transactions on Broadcasting*, 64(2):539–551, June 2018.

[36] Niall Murray, Oluwakemi A Ademoye, Gheorghita Ghinea, and Gabriel-Miro Muntean. A tutorial for olfaction-based multisensorial media application design and evaluation. *ACM Computing Surveys (CSUR)*, 50(5):67, 2017.

[37] Niall Murray, Brian Lee, Yuansong Qiao, and Gabriel-Miro Muntean. Multiple-scent enhanced multimedia synchronization. *ACM Trans. Multimedia Comput. Commun. Appl.*, 11(1s):12:1–12:28, October 2014.

[38] Niall Murray, Yuansong Qiao, Brian Lee, AK Karunakar, and Gabriel-Miro Muntean. Subjective evaluation of olfactory and visual media synchronization. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 162–171. ACM, 2013.

[39] Niall Murray, Yuansong Qiao, Brian Lee, Gabriel-Miro Muntean, and AK Karunakar. Age and gender influence on perceived olfactory & visual media synchronization. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[40] Marianna Obrist, Elia Gatti, Emanuela Maggioni, Chi Thanh Vi, and Carlos Velasco. Multisensory experiences in hci. *IEEE MultiMedia*, 24(2):9–13, 2017.

[41] N. Ranasinghe, K-Y. Lee, and E. Y-L. Do. Funrasa: an interactive drinking platform. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*, pages 133–136. ACM, 2014.

[42] N. Ranasinghe, T. N. T. Nguyen, Y. Liangkun, L-Y. Lin, D. Tolley, and E. Y-L. Do. Vocktail: A virtual cocktail for pairing digital taste, smell, and color sensations. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1139–1147. ACM, 2017.

[43] Nobuyuki Sakai, Sumio Imada, Sachiko Saito, Tatsu Kobayakawa, and Yuichi Deguchi. The effect of visual images on perception of odors. *Chemical Senses*, 30(suppl_1):i244–i245, 2005.

[44] J Simner and V Ludwig. What colour does that feel? cross-modal correspondences from touch to colour. In *Third International Conference of Synaesthesia and Art, Granada, Spain, April*, 2009.

[45] Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4):971–995, 2011.

[46] Charles Spence, Mary Kim Ngo, Bronwen Percival, and Barry Smith. Crossmodal correspondences: Assessing shape symbolism for cheese. *Food Quality and Preference*, 28(1):206 – 212, 2013.

[47] Naomi L Streeter and Theresa L White. Incongruent contextual information intrudes on short-term olfactory memory. *Chemosensory Perception*, 4(1-2):1–8, 2011.

[48] Benjamin Tag, Takuya Goto, Kouta Minamizawa, Ryan Mannschreck, Haruna Fushimi, and Kai Kunze. atmosphere: mindfulness over haptic-audio cross modal correspondence. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 289–292. ACM, 2017.

[49] Christian Timmerer, Markus Waltl, Benjamin Rainer, and Niall Murray. Sensory experience: Quality of experience beyond audio-visual. In *Quality of Experience*, pages 351–365. Springer, 2014.

[50] Augoustinos Tsiros. The parallels between the study of cross-modal correspondence and the design of cross-sensory mappings. In *Proceedings of the Conference on Electronic Visualisation and the Arts*, EVA '17, pages 175–182, Swindon, UK, 2017. BCS Learning & Development Ltd.

[51] Shafiq ur Réhman, Muhammad Sikandar Lal Khan, Liu Li, and Haibo Li. Vibrotactile tv for immersive experience. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–4. IEEE, 2014.

[52] M. Waltl, C. Timmerer, and H. Hellwagner. Improving the quality of multimedia experience through sensory effects. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 124–129, June 2010.

[53] J. Wang, B. Li, W. Hu, and O. Wu. Horror video scene recognition via multiple-instance learning. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1325–1328, May 2011.

[54] Jeffrey M Yau, Jonathon B Olenczak, John F Dammann, and Sliman J Bensmaia. Temporal frequency channels are linked across audition and touch. *Current Biology*, 19(7):561–566, 2009.

[55] Hyunho Yeo, Youngmok Jung, Jaehong Kim, Jinwoo Shin, and Dongsu Han. Neural adaptive content-aware internet video delivery. In *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*, pages 645–661, 2018.

[56] J. You. Attention driven visual qoe: Mechanism and methodologies. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 466–470, July 2013.

[57] Yuan Yuan, Chunlin Tian, and Xiaoqiang Lu. Auxiliary loss multimodal gru model in audio-visual speech recognition. *IEEE Access*, 6:5573–5583, 2018.

[58] Zhenhui Yuan, Ting Bi, Gabriel-Miro Muntean, and Gheorghita Ghinea. Perceived synchronization of mulsemedia services. *IEEE Transactions on Multimedia*, 17(7):957–966, 2015.

[59] Zhenhui Yuan, Shengyang Chen, Gheorghita Ghinea, and Gabriel-Miro Muntean. User quality of experience of mulsemedia applications. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):15, 2014.